White Paper

Report ID: 111503

Application Number: HJ-50187-14

Project Director: Elaine Treharne

Institution: Stanford University

Reporting Period: 1/1/2014-7/31/2016

Report Due: 10/31/2016

Date Submitted: 10/31/2016

National Endowment for the Humanities Project HJ-50187-14
Stanford Global Currents: PI Professor Elaine Treharne
White Paper

An international, inter-institutional research project, Global Currents, was funded in 2014 by the NEH, SSHRC, and NSERC among others. Spear-headed by Professor Andrew Piper at McGill University, the core team comprised Professor Mohammed Cheriet at ETS in Montreal, Professor Lambert Schomaker at Groningen, Professors Elaine Treharne and Mark Algee-Hewitt, and Dr Benjamin Albritton at Stanford University, and professors drawn from the McGill University scholarly community. Each core team received funding: ours at Stanford from the National Endowment for the Humanities' 'Digging into Data' Program began in February 2014.

The Aims of the Overall Project

Using textual corpora drawn from very different cultural contexts of production, the overall project has sought to combine new methods of image processing and machine learning with textual and codicological analyses in order to advance the understanding of the forms of global literary communication. Global literatures are a major focus of scholarly attention within current humanistic research, with the increasing ability to move beyond narrow textual expertises through the availability of large digitized corpora and the desire of scholars to work across traditional linguistic and national boundaries. To date, computational and data-impelled methods have not been regularly employed within humanities for this kind of work. The overall international project aimed to advance scholarly knowledge by simultaneously investigating two primary concerns (and here I quote from the original project proposal): firstly, "the challenges of non-western and non-print texts for computational analysis;" secondly, "the challenges of comparing vastly different forms of cultural expression from diverse periods and geographical locations."

As such, four groups were brought together to work on a diverse set of textual materials—in Latin, English, German and Chinese that ranged from *c.* 1080CE to *c.* 1900. The questions addressed of these texts centered on how digital tools can assist in ascertaining key features of informational design and transmission. These data—in forms of the manuscript codex, the printed book, and single-leaf or scroll technology—formed more than one million page images resulting in well over 100TB of processed data.

To answer this question over the two-year course of the grant, at Stanford, we used the following approach:

Visual Language Processing (VLP)

Thanks to Optical Character Recognition, large repositories containing hundreds of thousands of readable and interpretable digital texts exist for the world of printed books. No similar repository of handwritten textual objects is possible at present, because OCR itself does not work with the

variation inherent in script. The idiosyncrasies of chirographic technology, and the uniqueness of each medieval manuscript (or scroll, or single sheet document) are difficult to capture digitally, and, indeed, even in the most regular printed text, OCR itself has a significant error return rate.

This project has drawn on the expertise of Mohamed Cheriet's Synchromedia Lab and Lambert Schomaker's Institute for Artificial Intelligence and Cognitive Engineering, as they have developed different forms of Visual Language Processing (VLP) to determine similarities between folios (in Stanford's corpus): in the one case, similarities of lexical formation in handwritten materials; and in the other, relationships between elements that comprise the principal hierarchy of page layout (*mise-en-page*).

Our decision to use these methods means that for the first time, textual objects not usually analyzed by big-data approaches will be the focus of scholarly attention. Moreover, while large-scale digitization of manuscripts and other textual corpora is well underway, there are few tools available to permit cross-corpus visual investigations of writing and page layout. VLP brings new texts and new kinds of textual information to the forefront of extensive study of the literary past. Our contribution, focused on the visual features of historical, handwritten works allows an innovative, large-scale examination of scribal activities that are trans-chronological (in Stanford's case-study, dating from *c.*1080 to *c.*1220), multilingual (Latin and English, with occasional French), trans-regional, and multigeneric (histories, homilies, poetry, laws, prognostications, patristic tracts, scholastic texts, and so on).

Stanford Global Currents' Aims

Our project sought to examine the script and layout of a corpus of British manuscripts in English and Latin (with some French) from the post-Conquest period. These manuscripts represent a wide range of authors, regions, and genres. Our aim was to deduce if machine learning could identify particular information about handwriting and about retrieval tools from these folios. This would help us determine how manuscript producers manufactured their codices; how they assisted audiences in finding their way around the text; how the layout has changed or stayed the same across time and despite varied manuscript producers; and to understand the variation is *mise-en-page* features, particular between different languages.

Our corpus of manuscripts was supplied from the Parker Library of Old-English Manuscripts (http://parkerweb.stanford.edu/), a collaborative project between Stanford University and Corpus Christi College, Cambridge, consisting of 210 manuscripts dated between 1060 and 1220 with 63,000 total page images. Medievalists speak of the twelfth century as a major cultural turning point, one that witnessed the establishment of universities, scholastic analysis of texts, the growth of centralized governance, the institutionalization of the church, the codification of the law, the rediscovery of the classical past, and the period in Britain most characterized by a recorded and interpretable multilingual society. These manuscripts, which cover the three major literary languages in England and France, represent a profoundly rich resource for the study of how literacy and the recording of cultural memory came about on a grand scale. In particular, their

study using visual image processing might help tell us when the native vernacular tradition of book design and script ceded to that of the French colonizers in the post-Conquest period (from single column to double column layout, for example); how students (as opposed to clerics) began engaging with the written word and what differed about their processes of analysis (by discovering the emergence of complex scholastic manuscripts; by evaluating the role of commentaries and glosses); and what influence the universities had on the spread of literacy and education, particularly with regard to the production of manuscripts at particular centers (if we are able to deduce localization from VLP's corroborative evidence).

A secondary, but significant, research goal was to test the mechanism for large-scale image processing to be done on a corpus of digital resources held by an institutional repository in such a way that all new knowledge produced through analysis of those resources could be re-incorporated into the repository to enhance the digital resources themselves. This "virtuous circle" of scholarly communication, where a project consumes and then enriches re-usable repository data, has proven to be an ongoing challenge in the information sciences and library communities. Using the protocols specified by the International Image Interoperability Framework (IIIF)[1], the project provided images via API (rather than the more usual exchange of hard-drives through the post) and requested returned data be provided to conform to the IIIF specifications as well, insuring full re-usability of the results outside of the context of this particular project.

More specific, and object-focused research concerns were posited as:

A) Questions that focus principally on the ***mise-en-page* of the manuscript**, the page layout, and the ways in which scribes and manuscript compilers arranged text such that readers could negotiate the words and/or images (rarely does our corpus include images). Therefore:

1. Where is there notable white space on the folio?
2. What does that white space signify by way of textual or sub-textual denotation?

B) Other questions related to *mise-en-page* as less to do with space, and more to do with the **specific information retrieval tools** used by medieval scribes and designers. There are key features used regularly to denote hierarchies of text:

1. Rubrics

   Written in red ink, rubrics form, effectively, the title of the sermon or saint's life or chapter. This title is often little more than a pericope. Can rubrics be detected?

2. *Litterae Notabiliores*

   Litterae notabiliores are enlarged initials that are often decorated with simple flourishes and that act as visual cues for the beginning of a new textual item (very large initials) or a new 'paragraph' or section (smaller, pen-drawn initials). Can these be detected and somehow represented in the resulting data in terms of their proportional size?

3. Minor flourishes and decoration

[1] International Image Interoperability Framework: http://iiif.io

Many initials and individual graphs are decorated with red in-fill (tipped in using red ink) throughout the text. These often denote what we'd regard as new sentences. They're another means of finding one's way through the block of text. Can these be detected?

4. Marginal information

Catchwords (denoting the beginning of new quires) and run-overs, indicating that encapsulation of text that has run-over the manuscript line are interesting features. Can these be detected?

5. How are these elements of the *mise-en-page* linked conceptually and how will the program permit us to assess this?

C) Other spacing issues; issues of detail:

1. Inter-lexical spacing, inter-graphic spacing

Is it possible to detect the proportion of space between words or between individual letters? It is probable that this might be a major feature in determining scribal hands?

2. Is it possible to detect the shapes of particular graphs? For example, can the variation in the writing of the seven-shaped Tironian *notae* (for 'and') be noticed by the software?

Narrative of the Project and its Progress

Phase 1: Lexical Recognition Software

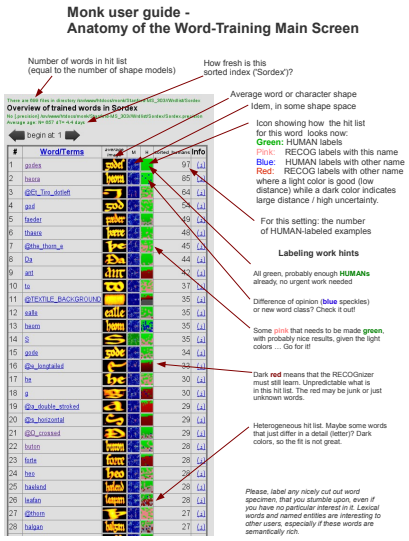
NEH 'Digging into Data' funding began in February 2014, and within a month, two undergraduate Research Assistants were appointed, to be joined by a third colleague in August, 2014. After preliminary discussion with the MONK project director, Lambert Schomaker, the students' work involved harvesting individual image files of medieval manuscripts from the Stanford Parker on the Web Repository to provide raw data for analysis. The RAs worked methodically through these hundreds of jpeg files to ensure that each image was as clean as possible, since image noise mars computer legibility. This stipulation immediately ruled out the provision of images with glosses and marginalia that we'd hoped to study, since any data-training of those features would be a separate process, outside of our scope in time and the ability of Schomaker's MONK project. Training the data in the way required proved an immensely time-consuming exercise and our return rate of clean images was 38% on the whole; that is, images of folios that were minus tears, repairs and holes; minus offsetting and show-through; minus dark patches and damage subsequent to manuscript manufacture; minus marginalia and interlinear glosses, annotations or *signes de renvoi*.

The images were selected in the first instance to illustrate the principal visual features that are most significant for determining information retrieval trends in *mise-en-page* design and page content, both in Latin and in English codices, from the long post-Conquest period. These visual features (listed below) were tagged by hand and the images and resulting data submitted to Professor Lambert Schomaker's projects, MONK and ALICE, at Groningen, and to Professor Mohamed Cheriet for his program at the École de technologie supérieure in Montreal. At the labs,

the manuscript images were processed using Visual Language Analysis and Network Modelling Tools, correlative to other textual corpora to begin to ascertain overarching elements of the texts' physical compositions across time and space.

Complex data was also provided beyond the level of the folio's characteristics to the MONK Lab at Groningen. This involved preparing images for lexical recognition software. For the first six months of the project, the RAs compiled successive wordlists of relatively high frequency complex words in Old English and in Latin to capture lexemes within the manuscript corpus to act as training data. This was a slow, and often frustrating, process, because these texts contain graphic characters that are not in contemporary use in English (or Latin); namely, ð, Ð, þ, æ, Þ, Æ, ꝑ, ꝓ, 7. So, for example, after submission of English words, such as *þe*, MONK would return it as *@the_thorn_e*, affecting legibility and slowing down the process. The program itself was also not entirely intuitive, altering orthography, and creating non-existent lexemes that required weeding out. The program mislabeled vernacular lexemes as Latin, and caused significant error that required individual manual correction. Even so, this was clearly creative failure in a sense: it was new research, building criteria that either effect analysis or that hinder it.

In determining the wordlists, in Old English, for example, the students tagged *gewrit, raedan, leornian, reordian, singan, sacerd, preost, bisceop*; in Latin wordlists, the students picked out lexemes, such as *liber, carta, codex, kalendarium, volumen, evangeliarum, litteratura, sermo, dictus*. The general themes of the wordlists were vocabulary items related to the church and education, so that we could track these words through a large corpus of manuscripts of homilies, patristic writings, and liturgical materials, once the program's efficacy had been proven. As with the initial folio selection, identifying the words within their manuscript context, and manually tagging them inside the MONK program, then searching individually again in other manuscript folios, and then manually correcting the words once the initial data was confirmed by MONK was a difficult process. This work took up to two hours per folio, though while immensely time-consuming, we assumed some promising consequences once fully automated. This was the training document used to begin the students' MONK encounter:



Each document was ingested by the software, but numerous glitches caused the preparation time to be very slow per folio, with more time required to iron out problems. Chief among these was that the software initially showed few signs of learning: even at the end of processing one of the manuscripts, when nearly sixty examples of the word "thing" had been entered, the software was still not able to reliably recognize the token. Secondly, it was difficult to manually set the zones for words, and at times there was no zone that perfectly sectioned off that word. Thirdly, a standardized list for Old English graphs had to be developed, together with instructions on how to enter these graphs into the computer. Fourthly, when a word search was manually entered, not all of the results were discoverable. Other efforts at word-training indicated that after tagging and entering a word across the manuscript in MONK, not every instance of that word was retrievable with the search function. This revealed the question of whether there might be any system to the white space being included in the training of orthographic words? Should white space be included before and/or after a word? This question had important implications for our work with the lab of Mohammed Cheriet, which focused principally on *mise-en-page*.

In this way, through the manual checking and correcting, the data recognized by MONK was affirmed or denied. According to Professor Schomaker, the harvested data was yielding excellent results, though there remain many design issues and other glitches to iron out. We are still awaiting complete results for fuller analysis, and for ascertaining accuracy, potential, and actual use in relation to scribal and orthographic profiling.

What did emerge unexpectedly from the harvesting of lexical data and the manual correction of Monk's identifications is the use to which the program might be put in future iterations. Through the collation of lexemes intertextually, scribal variation is easily visualizable. The distinctions in scribal habits between texts, between codices, and between scribal stints is made absolutely apparent; thus, for instance, the manner of abbreviating *nomina sacra*, often thought to be a relatively stable scribal practice, varies notably within individual manuscripts and

between manuscripts. When the program is fully tested and operable, it seems probable that important observations about trends, dates, and attribution can be made. We'll be looking for evidence for these sets of potential conclusions when the data is returned for analysis.

Phase 2: Visual Language Processing

Working towards the second part of data harvesting and provision in 'Global Currents', the Stanford team gathered information about the *mise-en-page* of twelfth-century Latin and English manuscripts. A large team of RAs moved through the textual corpus identifying multiple information retrieval tools (as below); they labeled them, and supplied large numbers of images as training data for Professor Cheriet's ETS lab in Montreal. Our examples were used by members of the ETS lab to train a classification algorithm that was able to recognize and extract the appearances of each feature throughout the corpus as a whole. The principal research questions asked how stable the *mise-en-page* features might be, irrespective of codicological form, generic function, chronology, or language of text. These features were, principally, running-headers, catchwords, writing grid format, *litterae notabiliores*, enlarged initials, minor flourishes and decorative devices, rubrics, intertextual space, ink-filled graphemes, and interlexical space. These are discussed in greater detail on our website: https://sites.stanford.edu/globalcurrents/discovery/visual-hierarchy, which is currently undergoing thorough proof-reading. The initial set of training data was numerically labeled by the RAs to include the four most important features for the program to identify; these were *litterae notabiliores*, enlarged initials, rubrics and intertextual space. The anticipation was that the RAs would check and verify computerized output, permitting full analysis of the harvested data in the light of the questions asked.

These sample identifiers were sent, together with tens of thousands of other images in the corpus, to ETS, and they developed and fine-tuned their program to identify these features throughout the corpus. The mechanism for transmission with ETS, a list of IIIF-compliant URLs (one for each image of a manuscript folio), allowed the Montreal team to write a simple script to harvest the images in a uniform way rather than having to go through the time-consuming process of copying and shipping the images on physical media. We had an excellent, close-working relationship with this team, and the benefits of clarity and focus were obvious, as our initial test cases yielded impressive results for analysis. At this point, early in the process, we had a batch of initial results returned, which employed the IIIF URL pattern (which allows a region of interest on an image to be expressed as coordinates in the URL), which Dr Benjamin Albritton was able to display through simple html galleries that gathered all of the results for a specific feature in a web page for visual perusal. The example image below illustrates a tiny portion of a small percentage of the overall image set, and it shows rows of *litterae notabiliores*.

This initial data contained a few algorithmic glitches, where some of the individual features were partially abbreviated or erroneously identified by the program. Our feedback to the team, mediated by our Project Manager, assisted the ETS team in fine-tuning the discovery algorithm for all the rest of the data and our results are outlined below.

At the same time as students worked their way through the training data compilation, they recorded their responses to the manuscript corpus. This was a particular component of the research overall (the interpretation of manuscript materials), which reflects the initial response to medieval materials by users, some of whom have never encountered this material before. This more reflective data is being scrutinized for any generalizations about audience response to digital images that might be extrapolated. The team at Stanford will determine if these initial audience responses can be employed in the design of better interpretative frameworks for digital repositories that present complex early textual materials, often to interested viewers who have little or no expertise in palaeography, codicology, and modern methods of curation and display.

Phase 2: Research Questions

Initial research questions concerned with *mise-en-page* were, it transpired during the second phase of the project, both deductive and inductive. First, deductive questions centered on the theory that we should be able to date manuscripts from trends discernible in the evolution of the major information retrieval tools we identified. Palaeographical and codicological developments in the second half of the twelfth century are critical, and include notable shifts in the complexity of folio design (double- or triple-column from single; introduction of running heads; systematization of rubrication; introduction of more navigational aids, including capitals, *capitula*; and recognition of the significance of clearly demarcated textual boundaries). We expected the data returned from ETS to permit a very rapid assessment of the accuracy of our forecasts, which were based on traditional, detailed scholarship, usually proceeding folio-by-folio, quire-by-quire, codex-by-codex.

The second deductive questions utilize the same features of page layout to ascertain if manuscripts can be localized to specific places of origin, based on similarity of feature. Localization remains one of the most vexed, but important, aspect of manuscript studies in modern scholarship: fewer than one-third of manuscripts can be assigned to a place of origin. We expected the rapid analysis of ETS's results, displayed through galleries, and annotated through the IIIF discovery environment, to provide a whole new perspective on what constitutes 'similarity' and 'difference' in manuscript production in the long twelfth century.

Inductive research questions leapt off the galleries put together by Dr Albritton from the raw data sent from Professor Cheriet's team. We were surprised to see how dissimilar particular *litterae notabiliores* are from others in the gallery. Dissimilarity might be attributable to national trends in color use; to the 'rusticity' of specific initials in manuscripts not produced at major writing centers; or to the idiosyncrasy of scribe-artists, who we might now be able to trace with greater precision. We were delighted to discover that manuscripts never before associated with one another might, in fact, be related in terms of their production methods. We saw this emerge through the serendipitous juxtaposition of initials in the gallery. Linked to manuscript images behind the thumbnails, we were swiftly able to recover all current scholarship on these books, and we are certain that we shall go on to break new ground in terms of geographical and chronological affinities between codices in this period as our research on these images matures.

Curation, Display and Evaluation

In the final stages of the project, the team worked on a number of different data analysis and display components that are discussed on our in-progress website (https://globalcurrents.stanford.edu/):

1. *Content development*

a) Population of Visual Hierarchy for each identified *mise-en-page* feature

i) Our resource uniquely displays *mise-en-page* features from different manuscripts with descriptive and some evaluative information. This is of great use to medieval scholars as well as interested general viewers. Our cogent resource also provides general definitions of these features and what they look like in context, with allied images for visual identification.

b) Compilation of descriptive overviews of the layout of each manuscript

i) The project team completed the initial sampling of manuscripts, from which we developed the training data for ETS in Montreal. Synoptic overviews of the *mise-en-page* of these manuscripts have been published.

ii) These overviews provide exceptional contextual material for our manuscript corpus in an easily accessible format.

2. Identification and notation of mise-en-page *features*:
On a folio-by-folio basis, RAs notated each occurrence of all features selected for analysis. These included overall layout, feature color, size, location, and aesthetic characteristics.

3. Metadata Provision
Initial results from ETS indicated that detailed metadata was essential to effect comprehensive investigation. As a result, new components of analysis were begun:

a) Development of a SQL database and querying tool to input results from ETS
The data received from the ETS lab came in a spreadsheet of IIIF URLs for full folios, and bounding box coordinates for detected features.

- ~1400 **Litterae Notabiliores** across 13 manuscripts (200 images)

https://stacks.stanford.edu/image/iiif/qk423pn9162%2F011_070_R_TC_46/full/pct:40/0/default.jpg	6-Litterae_Notabiliores	67	305	442	434
https://stacks.stanford.edu/image/iiif/qk423pn9162%2F011_083_V_TC_46/full/pct:40/0/default.jpg	6-Litterae_Notabiliores	606	2131	718	488
https://stacks.stanford.edu/image/iiif/qk423pn9162%2F011_085_V_TC_46/full/pct:40/0/default.jpg	6-Litterae_Notabiliores	1481	325	703	533
https://stacks.stanford.edu/image/iiif/qk423pn9162%2F011_099_V_TC_46/full/pct:40/0/default.jpg	6-Litterae_Notabiliores	1395	317	498	641
https://stacks.stanford.edu/image/iiif/qk423pn9162%2F011_099_V_TC_46/full/pct:40/0/default.jpg	6-Litterae_Notabiliores	1467	1281	441	440

- ~4,000 **Rubrics** across 17 manuscripts (800 images)

https://stacks.stanford.edu/image/iiif/zv088kx4487%2F003_075_V_TC_46/full/pct:40/0/default.jpg	7-Rubrics	1309	25	638	68
https://stacks.stanford.edu/image/iiif/ty948rv7120%2F009_11_R_TC_46/full/pct:40/0/default.jpg	7-Rubrics	1176	1286	757	72
https://stacks.stanford.edu/image/iiif/ty948rv7120%2F009_12_V_TC_46/full/pct:40/0/default.jpg	7-Rubrics	1560	352	759	72
https://stacks.stanford.edu/image/iiif/ty948rv7120%2F009_144_R_TC_46/full/pct:40/0/default.jpg	7-Rubrics	218	3495	710	100
https://stacks.stanford.edu/image/iiif/ty948rv7120%2F009_144_R_TC_46/full/pct:40/0/default.jpg	7-Rubrics	1088	3495	807	100

- ~9,800 **Enlarged Capitals** across 17 manuscripts (800 images)

https://stacks.stanford.edu/image/iiif/dt053nh0820%2F002III_273_V_TC_46/full/pct:40/0/default.jpg	9-Enlarged_Capitals	1471	592	323	158
https://stacks.stanford.edu/image/iiif/dt053nh0820%2F002III_274_R_TC_46/full/pct:40/0/default.jpg	9-Enlarged_Capitals	137	1619	46	156
https://stacks.stanford.edu/image/iiif/dt053nh0820%2F002III_274_R_TC_46/full/pct:40/0/default.jpg	9-Enlarged_Capitals	956	2444	58	489
https://stacks.stanford.edu/image/iiif/dt053nh0820%2F002III_282_R_TC_46/full/pct:40/0/default.jpg	9-Enlarged_Capitals	969	2350	183	169
https://stacks.stanford.edu/image/iiif/dt053nh0820%2F002III_286_R_TC_46/full/pct:40/0/default.jpg	9-Enlarged_Capitals	69	900	145	159

From this, we programmatically generated IIIF URLs for each feature and handed them off to students for tagging. For data storage, we employ a MySQL database. We chose SQL over a NoSQL database or simple JSON because, despite the recent surge in popularity of NoSQL systems like MongoDB, SQL is still the well-known standard and allows future growth of the dataset without sacrificing efficiency. To get a quick, preliminary look at the results while the database was in the works, we created a simple lazy-load/infinite-scroll HTML image gallery for each feature where every IIIF URL was hard-coded into an <img> tag. All of the scripting for IIIF manipulation, gallery creation, and SQL database loading was done in Python.

This database and tool provides the means to answer deductive research questions certainly facilitating the dating and localization of manuscripts.

b) Metadata attribution of results
RAs were tasked with exploring a sample of each detected feature from ETS' results and attributing metadata such as color and complexity, and, significantly, detecting errors in the feature modeling algorithms.

On the initial run-through with the first set of Enlarged Capitals (ECs) and *Litterae Notabiliores* (LNs), the RAs noted whether the image was captured in error (e.g. a marginal image instead of a LN) and what colors were present in the letter. The team noticed a fair amount of variability in terms of what was classified as a *Littera Notabilior*, and to a certain extent Enlarged Capitals. In order to capture some of this variability, the team decided on a classification system for LNs, which translates to the ECs as well. These categories were based directly on trends observed while evaluating the data, and essentially became a hierarchy of intricacy ranging from captured LNs that were a single-colored enlarged letter, typically at least three lines high, to multicolored and illuminated LNs that occupy the majority of a page and are often inhabited or historiated. The team settled on five distinct groupings for LNs, which include and range between these. One of the most common types was dual-colored LNs, with a larger capital letter typically in red or blue with flourished surrounding it in the other color.

- Initial Dataset excluding errors
- Sample of 10% of *each* feature from final dataset

URL	Letter	primary_color	secondary_colors	is_error	Error Description	ms_no	page	rv
https://stacks.stanford.edu/image/iiif/qk423	Q	red	blue			11	65	R
https://stacks.stanford.edu/image/iiif/qk423	U	red	blue			11	66	V
https://stacks.stanford.edu/image/iiif/qk423	O	red	blue			11	70	R
https://stacks.stanford.edu/image/iiif/qk423	U	blue	red			11	83	V
https://stacks.stanford.edu/image/iiif/qk423	U	red	blue			11	85	V
https://stacks.stanford.edu/image/iiif/qk423	P	red	blue			11	99	V
https://stacks.stanford.edu/image/iiif/qk423	C	blue	red			11	99	V
https://stacks.stanford.edu/image/iiif/qk423	C	red	blue	T		11	100	R
https://stacks.stanford.edu/image/iiif/qk423	V	blue	red			11	100	R
https://stacks.stanford.edu/image/iiif/qk423	N	red	blue			11	104	V
https://stacks.stanford.edu/image/iiif/qk423	T	red	blue			11	105	R
https://stacks.stanford.edu/image/iiif/qk423	E	blue	red			11	105	V
https://stacks.stanford.edu/image/iiif/qk423	S	blue	red			11	109	R
https://stacks.stanford.edu/image/iiif/qk423	T	blue	red			11	115	V
https://stacks.stanford.edu/image/iiif/qk423				T		11	115	V
https://stacks.stanford.edu/image/iiif/qk423	P	red	blue			11	116	R
https://stacks.stanford.edu/image/iiif/qk423	N	red	blue			11	118	R
https://stacks.stanford.edu/image/iiif/qk423				T		11	118	R
https://stacks.stanford.edu/image/iiif/qk423				T		11	118	R
https://stacks.stanford.edu/image/iiif/qk423	O	red	blue			11	119	V
https://stacks.stanford.edu/image/iiif/qk423	E	red	blue	T		11	144	V

In addition to categorizing the LNs, the researchers also typologized errors and marked those after noticing what was recurring. The most common of these occur when multiple correct LNs or ECs are captured in one bounding box, or in instances when other text or document noise is captures and misidentified as either a LN or EC. The researchers tagged the necessary subset of ECs and LNs, and worked through tagging intertextual space. They marked whether or not the image captured was an error; that is, something that clearly was not any type of intertextual space on the page, such as a full line of text. The team classified the errors into two types, textual and physical elements. For textual elements this was writing or drawing on the page, while examples of physical elements would be holes in the parchment. Thousands of these errors were tagged, while being scrupulous about maintaining accuracy.

4. Browsing and User Interface

Our lead undergraduate RA, Liz Fischer, created an IIIF gallery (similar to that created by Stanford's Dr Benjamin Albritton), to view the final results from ETS. The gallery can be found at https://stanford.edu/~efisch17/cgi-bin/globalcurrents/gallery/, but is currently not viewable by non-Stanford researchers, because the Parker on the Web images are not yet open access (they will be in the next eighteen months).

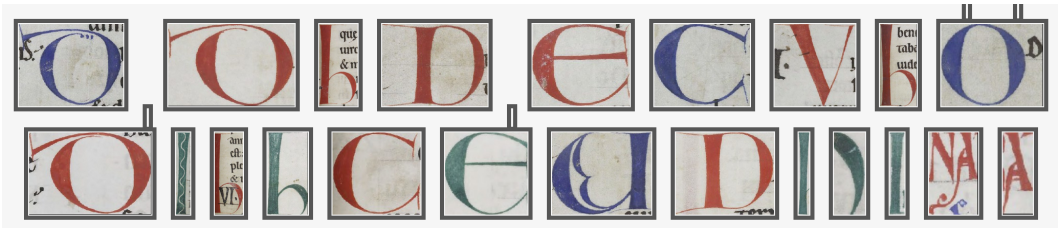
The gallery has had useful consequences in permitting the team to formulate and begin to answer globally significant research questions. For instance, from experience of working with medieval manuscripts, it might be assumed that green is a prevalent color in the embellishment of large capitals. Our results indicate that this is not the case, and that where green does occur, it may have important information to provide about date and place of origin of the manuscript. Our rapid overview of manuscript *mise-en-page*, facilitated by the gallery of images, also intimates that it is possible to offer a chronological typology of features of decoration; of the introduction of running headers; of the uses of rubrics; of the tendencies towards effects, like diminuendo display scripts, by particular scriptoria at particular times. Some of these research questions have emerged through obvious patterns visible in the data, and the subsequent close analysis of these apparent trends.

5. Numerical results from feature modeling

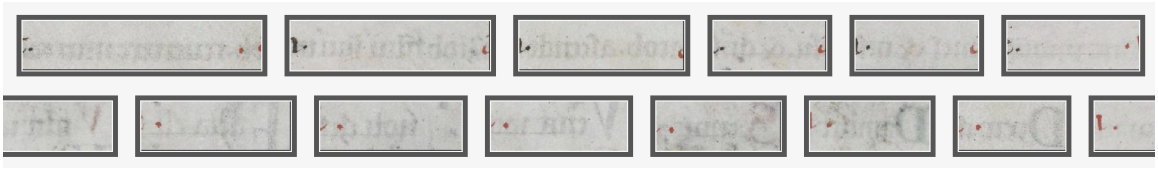
a) Having received our final dataset, the ETS computer software working with the training data we supplied was able to detect far greater numbers of information retrieval tools in manuscripts from *c.*1080CE to 1220CE than we had anticipated, and with far higher rates of error-free success than was thought possible:

i) 14,358 *Litterae Notabiliores*

ii) 66,291 Enlarged Capitals



iii) 247,861 Intertextual Spaces



iv) 171,438 Rubrics

Success Rates from ETS Feature Modeling Software

Stanford dataset

- Detection of 4 features

Feature	No. of Images	No. of samples	Precision	recall	f-measure
Litterae Notabiliores	195	231	0.95	0.61	0.74
Rubrics	864	5800	0.81	0.70	0.75
Intertextual Space	469	1500	0.60	0.75	0.67
Enlarged Capitals	776	3800	0.83	0.50	0.62

Findings for four key features of *mise-en-page*

There are key features used regularly to denote hierarchies of text. As we stated on page 3, the hierarchy of information retrieval tools selected included Rubrics, *Litterae notabiliores*, flourishes, enlarged capitals and intertextual space. Here, in red, are our conclusive results:

1. *Rubrics*

Written in red ink, rubrics form, effectively, the title of the text. Can rubrics be detected through our program? We have determined that rubrics can be detected using machine-learning algorithms identifying document layout analysis, color detection, and feature separation.

2. *Litterae Notabiliores*

Litterae notabiliores are enlarged initials that are often decorated might indicate the beginning of a new textual item or a new 'paragraph' or section. Can these be detected and represented in terms of their proportional size? *Litterae Notabiliores* can be detected using a fractal-based methodology, in addition to the detection methods used for rubrics.

3. Minor flourishes and decoration

Many initials and individual graphs are decorated with red in-fill. These often denote new phrases or sentences. Can these be detected? While these were not included in our main four features of analysis, based on our current results, it would be possible to detect these features.

4. Marginal information

Catchwords (denoting the beginning of new quires) and run-overs, indicating that encapsulation of text that has run-over the manuscript line, are interesting features. Can these be detected? While these were not included in our main four features, based on our current results, it would be possible to detect these features.

5. Conceptual Linking of mise-en-page elements?

We shall move forward by investigating the relationships between information retrieval tools by incorporating a sequence of entailments into our computational discovery.

III) Other Issues of Detail

1. Inter-lexical spacing, intergraphic spacing

Is it possible to detect the proportion of space between texts? The ETS lab has been able to successfully detect intertextual space. Within our corpus they found 247861 instances of this space. How this detection process might be finessed for intergraphic space is something to consider in ongoing research.

Obstacles to Progress

We encountered some serious obstacles to progress in the first year of the project. The first was the level of training required for research assistants was much greater than we had supposed. The relatively small budget supplied by the 'Digging into Data' grant means that best value for money is obtained by employing outstanding *undergraduate* research assistant. While intellectually exceptional, these RAs require training assistance, and flexibility for the completion of their coursework. This, coupled with the very time-consuming process of harvesting, training and correcting data for the MONK program, meant that progress was much slower than we had anticipated, with not a great deal of manipulable data to work with conclusively.

As such, towards the end of the first year, the PI decided it was necessary to submit a revised budget (not a request for additional support, but, rather, the employment of additional help to mange the data harvesting and training materials), supplied to the NEH in late February

2015. To quicken the pace of work, while maintaining the high quality of textual and material scrutiny, we increased our team of RAs, brought in an accomplished project manager (Celena Allen), and we relocated work to the Center for Spatial and Textual Analysis where the students could work collaboratively in an open office. The PI decided to focus much more energetically on the ETS' research sets, which were more promising than those supplied by MONK.

Future Goals

The project team is deeply impressed by the work of Professor Cheriet's lab at ETS in Montreal. As we make our way through tens of thousands of images that must be related back to their specific manuscript contexts, we are deciding upon the fundamental research questions we can go on to answer. We are also determining the new questions that have unexpectedly emerged from these path-breaking methodologies.

We are certain that this is important research, because the ETS program permits us to move swiftly through discrete components of manuscript production, insisting on new perspectives and fresh insight. Significant components of manuscripts have never before been seen in this way. Focusing on singular components aligned, often fortuitously, really does show this old material in a new light. We are already making significant discoveries, but more, our analysis of what the computer programs can do demonstrates the benefits of utilizing computational tools to answer deeply traditional humanistic questions. As such, this is important work for scholars in the medieval period, who are often entirely able to see the trees, but then fail to see the forest.

Conclusion and Outputs

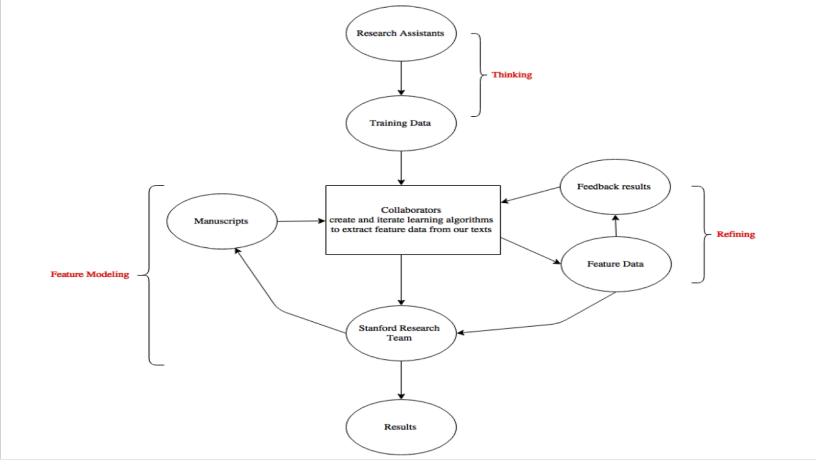
Different teams, with different literary corpora, were part of the efforts of the teams to determine how computer software can be trained to read specific textual features. Of the teams, Stanford's medieval manuscripts yielded the most reliable results for feature modeling discovery. Stanford's team previewed its data analysis at the IEEE International Conference in Santa Clara, and we have given numerous presentations to colleagues. We presented a project report at the 51st International Medieval Conference at Western Michigan University in May 2016; and we presented our final results in a very successful project session at the International Medieval Congress, University of Leeds in July 2016. The feedback we received from these presentations was tremendously helpful and overwhelmingly positive. Our undergraduate researchers were formal presenters at this session—something that truly impressed the audience.

Stanford Global Currents launched its website, https://globalcurrents.stanford.edu/, to wide interest and we are regularly updating and expanding the information provided on it. We shall be seeking an ISBN for this website to form an e-book. We have plans to link Global Currents' data with the data from EM1060to1220, a still-live AHRC-funded project that ran from 2005-2010 in the United Kingdom at the University of Leicester (http://www.le.ac.uk/english/em1060to1220/), and to continue working on the methodology and hierarchy of information retrieval tools in the coming two or three years. Moreover, the usable

data that we have amassed will be incorporated into future versions of the Parker on the Web project that will shortly become an Open Access resource (https://parker.stanford.edu/parker/actions/page.do?forward=home)

We anticipate completing a published paper with the team from ETS, and a full white paper with the whole international project team. Elaine Treharne is writing an academic article on the uses of computational and digital tools for manuscript studies, and she has completed preliminary detailed research on a number of features of *mise-en-page*, research that has directly emerged from scrutiny of the galleries of *Litterae Notabiliores* and Enlarged Capitals. New research findings (contingent upon the natures and genres of the manuscripts collected by Matthew Parker in his sixteenth-century antiquarian efforts) include interesting discoveries about the relationships between manuscripts that have never previously been linked; the rarity of particular letter forms; the scarce use of colors, such as purple and yellow; and the major distinctions between the production of vernacular manuscripts—where few expensive materials are employed in the production of text. This research has the capacity to make more nuanced scholars' understanding of how, when and where manuscripts were produced in England in the long twelfth century, and, particularly, how particular trends in production were trans-regional, even if resource and expertise differed from place to place.

Workflow Schema



Report Compiled by Elaine Treharne and with contributions by Celena Allen, Benjamin Albritton, Mark Algee-Hewitt, Matt Aiello, Liz Fischer and Clare Tandy

Full Project Team

- Professor Andrew Piper, McGill University, Montreal: Inter-institutional PI
- Professor Elaine Treharne, Stanford University, Stanford: Stanford PI
- Professor Mohammed Cheriet, ETS, Montreal: Software Developer and Consultant

- Professor Lambert Schomaker, Groningen, Netherlands: Software Developer (MONK) and Consultant to March 2015 (see Project Reports 1-3)
- Professor Mark Algee-Hewitt, Stanford University, Stanford: Co-Director
- Dr Benjamin Albritton, DLSS, Stanford University Libraries, Stanford: Technical Consultant and Co-Director
- Celena Allen, CESTA, Stanford University: Project Manager
-

Stanford Researchers

Lead RAs 2014-2016: Matthew Aiello, Liz Fischer, Clare Tandy

RAs 2014-2016: Benjamin Diego, Rukma Sen, Clare Tandy, Andrew Lee, Ryan Smith, Brooke Mandujano, Jaye Boissiere, Madelaine Bixler, Jasmine Guillory, Jaclyn Marcatili, Lindsay Mewes, Aisha Sharif